Predicting the Function of Hypothetical Protein PANDA_003700 using Computational Analysis Methods

Cameron Bixby and Padmanabhan Mahadevan*

Dept. of Biology, University of Tampa, Tampa, FL 33606
*To whom correspondence should be addressed (pmahadevan@ut.edu)

Abstract

The majority of the gene products produced after an organism is sequenced are proteins whose function is not known, called hypothetical proteins (HPs). Proteins that are predicted from nucleic acid sequences only and proteins with unknown functions are considered hypothetical proteins (Lubec et al., 2005). Therefore as large amounts of hypothetical proteins are discovered from genomic sequencing, they will continue to enter the spotlight of many studies in the Bioinformatics and Genomics field. About half of the proteins in most genomes are candidates for HPs (Lubec et al., 2005). Therefore, determining the function of the HPs is very important when trying to complete the genomic and proteomic information of a sequenced organism. HPs are observed across a variety of phylogenetic lineages but their functions are not characterized (Galperin & Koonin, 2004). Therefore, the challenge to characterize the function of HPs using experimental and computational methods has become more important in genomic studies.

Typically, the work dedicated to discovering the functions of HPs can be separated into two parts: prediction of protein function through its sequence and prediction of the 3-D structure of an HP. In terms of predicting the function of a HP through its sequence, researchers will use computational methods in order to compare their HP against functional proteins in hope of high sequence identities. In finding the similarities between sequences, researchers can infer the function of the protein, explore protein families, and evolutionary relationships (Lubec et al., 2005). The most common tool in calculating sequence similarity is the Basic Local Alignment Search Tool (BLAST) which has a version that can blast a protein query against a database of proteins. Exploration of various protein families to see if the HP shares any common evolutionary origin is another route taken by researchers in order to gather more information on their HP. Protein families are sets of protein regions which share a significant degree of sequence similarity (Punta et al, 2011). Therefore using databases like Pfam can display various relationships between a HP and other functional proteins. It is also important to mention that protein domains are also considered another area in which a researcher can use the HP sequence to discover its domains. Protein domains are viewed as the basic components of proteins and from this it helps determine the functional characterization (Veretnik et al., 2004).

The sequence of a hypothetical protein can provide a lot of insight in terms of the prediction of the protein structure itself which can then further help determine the function of the HP. This ties in with the goal of structural genomics which is to create a complete inventory of protein folds/structures that can help predict functions for all proteins (Mittl & Gütter, 2001). One way to determine the 3-D structure of a protein is by...
either x-ray crystallography or NMR, then the structure can be compared against other structures in a protein database (Zarembinski et al., 1998). However, those experimental methods are usually difficult, complex, and time consuming. Therefore, with limited experimental models of proteins to compare with a HP, homology modeling has become a reliable way to determine the 3-D structure by using a HP’s amino acid sequence. It is important to mention that homology searches are more accurate if the sequence similarity between the HP and the homolog of another known protein is greater than 30%. Overall, stronger the sequence similarity of a HP to other functional proteins, the likelihood of predicting its structure and function increases tremendously.

Computational tools have allowed researchers to generate more information about HPs in sequenced genomes across various organisms such as, mammals. Hypothetical proteins constitute a large portion of mammalian proteomes (Lubec et al., 2005) which can possibly reveal inferred evolutionary relationships between other mammals allowing for comparison. Therefore, a hypothetical protein was chosen at random from the sequenced genome of the Giant Panda (*Ailuropoda melanoleuca*). The HP chosen was PANDA_003700 and in an effort to gain insight into the process of determining the function of this HP, various computational analysis methods were implemented in the study. It was hypothesized that the function of the HP from the Giant Panda could be determined with the use of various genomic computational analysis methods. After subsequent analysis, the predicted function of the HP PANDA_003700 was that of an Mrt4 protein which is involved in ribosomal biogenesis.

References


